DOCUMENT RESUME

ED 280 896 TM 870 228

AUTHOR O'Brien, Francis J., Jr.
TITLE A Derivation of the Unbiased Standard Error of Estimate: The General Case.
PUB DATE 87
NOTE 54p.; For earlier monographs in this series, see ED 215 894, ED 216 874, ED 223 429, and ED 235 205.
PUB TYPE Reports - Research/Technical (143) -- Guides - Classroom Use - Materials (For Learner) (051)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Error of Measurement; *Estimation (Mathematics); Goodness of Fit; Higher Education; *Mathematical Models; *Predictor Variables; Proof (Mathematics); *Raw Scores; Regression (Statistics); Statistical Studies
IDENTIFIERS Applied Statistics; *Z Scores

ABSTRACT

This paper is part of a series of applied statistics monographs intended to provide supplementary reading for applied statistics students. In the present paper, derivations of the unbiased standard error of estimate for both the raw score and standard score linear models are presented. The derivations for raw score linear models are presented in graduated steps of generality for one, two, three, and any finite number of predictors. A brief overview of regression analysis precedes the derivations. Appendices include: (1) errata for a derivation of the sample multiple correlation formula; and (2) a discussion of linear and nonlinear regression models. (LMO)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED280896

A Derivation of the Unbiased Standard Error of Estimate: the General Case

Francis J. O'Brien, Jr., Ph.D.

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

F. J. O'Brien, Jr.

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

© 1987
Francis J. O'Brien, Jr.

ALL RIGHTS RESERVED

BEST COPY AVAILABLE

TM 870 228

ERRATA SHEET for "A Derivation of the Unbiased Standard Error of Estimate: the General Case"

PAGE	NOW READS	CHANGE TO
4, 2nd equation	$(X - X_1)$	$(X - \bar{X}_1))$
6, definition of R^2	$R^2_{Y.x_1,x_2,\ldots,x_j,\ldots x_p}$	$R^2_{Y.x_1,x_2,\ldots,x_j,\ldots,x_p}$
9, definition of r_{y1}	$\sum x_1 Y$	$\sum x_1 y$
9, footnote b	$\sum X_1 Y$	$\sum X_1 y$
9, definition of x_1	--	align subscript for $\bar{X}$
10, 1st equation	$S_{Y.x1}$	$S_{Y.x_1}$
12, 5th equation	$\sum x_1 Y$	$\sum x_1 y$
18, 2nd equation	$(b_1^2 S_1^2 + b_2^2 S_2^2 + 2b_1 b_2 r_{y1} S_1 S_2)$	$(b_1^2 S_1^2 + b_2^2 S_2^2 + 2b_1 b_2 r_{12} S_1 S_2)$
26, 3rd line from bottom	$-2(\sum_{j=1}^{3} b_j r_{yj} S_j S_Y)$	$-2(\sum_{j=1}^{3} b_j r_{yj} S_j S_y)$

PAGE*	NOW READS	CORRECT TO
31, 2nd line of 2nd equation	$b_1 \sum^{2} x_1$	$b_1 \sum^{2} x_1^{2}$
34	$\sum_{j=1}^{p} b_j r_{yj} S_Y S_j$	$\sum_{j=1}^{p} b_j r_{yj} S_y S_j$

*Refers to page at top.

Table of Contents

	Page
Introduction	1
Overview of Derivation	1
Overview of Regression Analysis	2
The Standard Error of Estimate	4
Derivations for Raw Score Model	8
Derivation for One Predictor	8
Derivation for Two Predictors	14
Derivation for Three Predictors	20
Derivation for p Predictors	29
Derivations for Standard Score Model	36
Introduction	36
Derivation for One Predictor	37
Outline for Derivations	40
Appendix A:Errata for ED 235 205	41
Appendix B:Discussion of Linear and Nonlinear Regression Models	42
Notes	46
References	48

List of Tables

Table		Page
1.	Basic Sample Descriptive Statistics for One Predictor Raw Score Model	9
2.	Substitution Equations for Two Predictor Raw Score Model	17
3.	Functions of R^2 for Two Predictor Raw Score Model	19
4.	Generalized Substitution Equations for Raw Score Model	25
5.	Functions of R^2 for Three Predictor Raw Score Model	27
6.	Functions of R^2 for p Predictor Raw Score Model	34

A Derivation of the Unbiased Standard Error of Estimate: the General Case

Francis J. O'Brien, Jr., Ph.D.

Introduction

This paper represents the fifth in a series of applied statistics monographs (See O'Brien 1982a, 1982b, 1982c, 1983a). The purpose of these papers is to provide supplementary reading for applied statistics students. The intended audience is social science graduate and advanced undergraduate students. The minimum background for most of the existing and forthcoming papers is familiarity with elementary analysis of variance, and multiple correlation and regression analysis.

The unique feature of this series is detailed proofs and derivations of important formulas and relationships which are not readily available in textbooks, journal articles and similar sources. Each proof or derivation is presented in a detailed and clear fashion using well defined and consistent notation. When necessary, a review of relevant algebra is provided. Calculus is not used or assumed.

The present paper assumes familiarity with two previous papers in this series (O'Brien, 1982c, 1983a). Each paper formulated a detailed derivation of the multiple correlation formula of one criterion and p predictors for the linear model. The first paper (1982c) presented a derivation of the multiple R based on standard (Z) scores,[1] and the second showed the analogous derivation for the raw score model.[2]

Overview of Derivation

In the present paper derivations of the unbiased standard error of estimate for both the raw score and standard score linear models are presented. The derivations will be presented in graduated steps of generality. First the derivation for one criterion (dependent) variable and one predictor (independent) variable is presentd for the raw score model. A derivation for two raw score predictors is then presented. Next, the derivation for the three predictor case is formulated. Finally, the derivation for any (finite) number of predictors is presented. Derivations for the standard score model are then outlined.

Overview of Regression Analysis

Prior to presenting the derivations, a brief overview[3] of regression analysis will be given. Let us consider the linear regression model for one raw score criterion and one predictor. Assume one is attempting to predict one criterion with one predictor. We assume that the model[4] is linear in form. The mathematical model we might select to "fit" such a distribution is the simple linear equation:

$$\hat{Y} = a + b_1 X_1$$

Where:

$\hat{Y}$ = the predicted criterion,

a = the slope intercept term,

b_1 = the slope coefficient term,

x_1 = the predictor variable in deviation score form ; i.e.,

$$x_1 = X_1 - \overline{X}_1 \quad \text{where} \quad \overline{X}_1 \text{ is the arithmetic mean.}$$

If a scatter diagram were constructed for this hypothetical model (based on actual data, of course), the actual raw score observations would in all likelihood not fall on the line defined by the linear equation of the idealized mathematical model ($\hat{Y}$).

Such deviations from $\hat{Y}$ are considered errors of prediction. We can conceive a raw score observation as consisting of a component predicted by the model plus an error component. That is:

$$Y = \hat{Y} + e$$

Where:

Y = the actual criterion we want to predict by $\hat{Y}$,
e = the amount of numerical error resulting from using the idealized mathematical model $(\hat{Y})$ to predict the actual raw score criterion (Y).

That is, an actual dependent (criterion) variable score consists of the quantity predicted by the idealized "best fitting" line plus an error component.

The error made in predicting the observed criterion score by the model is simply:

$$e = Y - \hat{Y}$$

One of the goals of regression analysis is to minimize the prediction error denoted by e above. It can be seen that if e=0, then the actual criterion is perfectly predicted by the selected mathematical model. That is to say, the simple linear

equation fitted to the observed data points, $a + b_1 x_1$,
predicts every observation (Y) in the distribution. Geometrically,

when e=0, every Y score falls on the straight line, $\hat{Y}$. For this case, the values corresponding to a and b can be solved empirically using elementary algebra based on the observed data. Rarely, however, do such distributions exist in the social sciences. Consequently, we are forced to select procedures which will provide computing formulas for calculating the a and b terms.

The technique most often used in the social sciences to minimize the error of prediction is the "least squares" procedure. Essentially, this procedure seeks to maximize predictability by minimizing prediction error. The least squares criterion or goal is summarized in the following
[5]
expression:

$$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2 = \text{a minimum}$$

If we substitute the quantity for $\hat{Y}$ previously defined, we can rewrite the least squares criterion as:

$$\sum \left[Y - (a + b_1 x_1) \right]^2 = \sum (Y - a - b_1 x_1)^2 = \sum e^2 = \text{minimum}$$

(As an aside, "least squares" means we determine values for a and b such that the squared error term results in the least possible value).

The Standard Error of Estimate

The standard error of estimate provides a measure of the average amount of error that results from using $\hat{Y}$ for Y score prediction. (See Lindeman, et al.). The unbiased standard error of estimate for one predictor is defined as follows:

$$S_{Y.x_1} = \sqrt{\frac{\sum (Y - \hat{Y})^2}{n-2}} = \sqrt{\frac{\sum \left[Y - (a + b_1 (X_1 - \bar{X}_1) \right]^2}{n-2}}$$

$$= \sqrt{\frac{\sum (Y - a - b_1 x_1)^2}{n-2}}$$

where:

$S_{Y.x_1}$ = the unbiased standard error of estimate for one predictor,

n = the sample size.

Note that the predictor variable (x_1) is in deviation form. However, the criterion to be predicted (Y) is not transformed; nor do we transform the predicted criterion ($\hat{Y}$).

This is the definitional formula for the unbiased standard error of estimate. An equivalent formula shown in virtually all applied statistics textbooks is as follows:

$$S_{Y.x_1} = S_y \sqrt{\frac{n-1}{n-2}\left(1-r^2_{x_1 y}\right)}$$

where:

S_y = the standard deviation of the actual criterion score,

$r^2_{x_1 Y}$ = the square of the simple Pearson correlation between the predictor in deviation form (x_1) and the criterion (Y).

This formula will be derived in this paper.

In general, the standard error of estimate can be obtained for a linear regression model containing any finite number of predictors. If we let p represent an indefinite number of raw score predictors, the unbiased standard error of estimate can be expressed as::

BEST COPY AVAILABLE

$$S_{Y.x_1,x_2,\ldots,x_j,\ldots,x_p} = \sqrt{\frac{\sum(Y-\hat{Y})^2}{n-(p+1)}}$$

$$= S_y \sqrt{\frac{(n-1)\,(1-R^2_{Y.x_1,x_2,\ldots,x_j,\ldots,x_p})}{n-(p+1)}}$$

where:

$S_{Y.x_1,x_2,\ldots,x_j,\ldots,x_p}$ = the unbiased standard error of estimate for p predictors (in deviation score form),

p = an indefinite number of predictors,

$R^2_{Y.x_1,x_2,\ldots,x_j,\ldots,x_p}$ = the squared linear multiple correlation between one criterion and p predictors.

This formula also will be derived in this paper.[6]

The standard error of estimate also can be derived for regression models in which the variables have been expressed in standard score (Z) form. The unbiased sample standard error of estimate for a one predictor standard score linear

model is defined as :[7]

$$S_{Z_Y.Z_1} = \sqrt{\frac{\sum(Z_Y - Z_{\hat{Y}})^2}{n-2}} = \sqrt{\frac{\sum\left[Z_Y - (A + B_1 Z_1)\right]^2}{n-2}}$$

where:

$S_{Z_Y.Z_1}$ = the standard error of estimate for the standardized criterion (Z_Y) and the standardized predictor (Z_1),

n = the sample size,

A = the slope intercept term,

Z_1 = the standardized predictor,

B_1 = the beta (regression) weight.

e_Z = the prediction error.

We show that the definitional formula above is equal to:

$$S_{Z_Y.Z_1} = \sqrt{\frac{n-1}{n-2}(1-r^2_{Z_Y,Z_1})}$$

where: $r^2_{Z_Y, Z_1}$ = the squared correlation of Z_Y and Z_1.

For standard score variables, the unbiased standard error of estimate for p predictors is:

$$S_{Z_Y.Z_1,Z_2,\ldots,Z_j,\ldots Z_p} = \sqrt{\frac{n-1}{n-(p+1)}\left[1 - R^2_{Z_Y.Z_1,Z_2,\ldots,Z_j,\ldots,Z_p}\right]}$$

where:

$S_{Z_Y.Z_1,Z_2,\ldots,Z_j,\ldots,Z_p}$ = unbiased standard error of estimate for p predictors,

$R^2_{Z_Y.Z_1,Z_2,\ldots,Z_j,\ldots,Z_p}$ = squared multiple correlation between the criterion (Z_Y) and p standardized predictors.

In this paper we will concentrate on the standard error of estimate for the raw score model. The derivations for the Z score model will be outlined. The reader may wish to work out the derivations for the standard score model using the detailed presentations for the raw score model as a guide.

Derivations for Raw Score Model

In the next several sections, we will show the derivations of the unbiased standard error of estimate for raw scores. We begin with the simplest case of one criterion and one predictor.

Derivation for One Predictor

For the readers convenience in working through the algebra, we will summarize relevant definitions and formulas. This is done in Table 1.

Table 1

Basic Sample Descriptive Statistics for One Predictor
Raw Score Model

Regression Model [a]:	$\hat{Y} = a + b_1 x_1 = \bar{Y} + r_{y1} \frac{S_y}{S_1} x_1$
Variance of Y:	$S_y^2 = \frac{\sum (Y-\bar{Y})^2}{n-1} = \frac{\sum y^2}{n-1}$
Variance of X_1:	$S_{X_1}^2 = \frac{\sum (X_1-\bar{X}_1)^2}{n-1} = \frac{\sum x_1^2}{n-1}$
Correlation of x_1 and y: [b]	$r_{y1} = \frac{\sum x_1 Y}{(n-1) S_Y S_1}$

Note: All summations range from i=1 to i=n observations.

[a] This is derived from the least squares criterion ; i.e.,

$$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - a - b_1 x_{1i})^2 = \sum_{i=1}^{n} e_i^2 = \text{minimum}$$

See O'Brien, 1983a, p. 44

[b] See O'Brien, 1983a, for justification that the numerator in the correlation formula may be given as:

$$\sum x_1 Y, \quad \sum x_1 y \text{ or } \sum X_1 Y, \text{ where } x_1 = X_1 - \bar{X}_1 \text{ and } y = Y - \bar{Y}.$$

In this paper, we will use the correlation expression

r_{y1} (or r_{y2}).

We begin by repeating the definition of the unbiased standard error of estimate:[8]

$$S_{Y.x1} = \sqrt{\frac{\sum (Y - \hat{Y})^2}{n-2}}$$

Substituting for $\hat{Y}$:

$$S_{Y.x_1} = \sqrt{\frac{\sum (Y - a - b_1 x_1)^2}{n-2}}$$

It will be easier if we work with the variance error of estimate. This is simply the square of the standard error of estimate:

$$S^2_{Y.x_1} = \frac{\sum (Y - a - b_1 x_1)^2}{n-2}$$

It was shown by the author that the slope intercept term, a, is equal to the criterion mean, $\bar{Y}$ (See O'Brien, 1983a, p.44). Making that substitution and rearranging terms:

$$S^2_{Y.x_1} = \frac{\sum \left[(Y-\bar{Y}) - b_1 x_1 \right]^2}{n-2}$$

Let us express $(Y-\bar{Y})$ in deviation score form to simplify the algebra: $y = Y-\bar{Y}$. This gives us:

$$S^2_{Y.x_1} = \frac{\sum (y-b_1 x_1)^2}{n-2}$$

Squaring out the terms inside parentheses for this binomial expression:

$$S^2_{Y.x_1} = \frac{\sum (y^2 + b_1^2 x_1^2 - 2yb_1 x_1)}{n-2}$$

Bringing the summation operator inside and factoring constants outside the summation operator (recall that b_1 functions as constant to be estimated in the regression model):

$$S^2_{Y.x_1} = \frac{\left(\sum y^2 + b_1^2 \sum x_1^2 - 2b_1 \sum x_1 y \right)}{n-2}$$

Substituting the following expressions (see Table 1):

$$\sum y^2 = (n-1)\, S_y^2$$

$$\sum x_1^2 = (n-1) S_1^2$$

$$b_1 = r_{y1} \frac{S_y}{S_1}$$ (based on substitution from Table 1 and O'Brien, 1983a, p.44)

$$\sum x_1 Y = (n-1) r_{y1}\, S_y\, S_1$$

Thus:

$$S^2_{Y.x} = \frac{1}{n-2} \left[(n-1) S_y^2 + (r_{y1}\, S_y / S_1)\, (n-1) S_1^2 - 2(r_{y1}\, S_y / S_1)\, (n-1) r_{y1}\, S_y\, S_1 \right]$$

Factoring out the (n-1) term:

$$S^2_{Y.x_1} = \frac{(n-1)}{(n-2)}\left[S^2_y + r^2_{y1}(S^2_y/S^2_1)S^2_1 - 2r^2_{y1}S^2_y\right]$$

Simplifying:

$$S^2_{Y.x_1} = \frac{(n-1)}{(n-2)}\left[S^2_y + r^2_{y1}S^2_y - 2r^2_{y1}S^2_y\right]$$

$$= \frac{(n-1)}{(n-2)}\left[S^2_y - r^2_{y1}S^2_y\right]$$

$$= S^2_y\frac{(n-1)}{(n-2)}\left[1-r^2_{y1}\right]$$

Taking the (positive) square root, the unbiased standard error of estimate for one raw score predictor is:

$$S_{Y.x_1} = S_y\sqrt{\frac{(n-1)}{(n-2)}\left[1 - r^2_{y1}\right]}$$

END OF PROOF

Derivation for Two Predictors

In this section we seek to show that the unbiased standard error of estimate for two raw score predictors is:

$$S_{Y.x_1,x_2} = S_y \sqrt{\frac{(n-1)}{(n-3)}\left[1 - R^2_{Y.x_1,x_2}\right]}$$

where:

S_y = the observed criterion standard deviation,

$R^2_{Y.x_1x_2}$ = the squared multiple correlation between the criterion and the two raw score predictors (in deviation score form)

We begin with the definition of the unbiased standard error of estimate for two raw score predictors:

$$S_{Y.x_1,x_2} = \sqrt{\frac{\sum(Y-\hat{Y})^2}{n-(p+1)}} \cdot \sqrt{\frac{\sum(Y-\hat{Y})^2}{n-3}} \cdot \sqrt{\frac{\sum(Y-a-b_1x_1-b_2x_2)^2}{n-3}}$$

As in the one predictor derivation, it will be easier to work with the variance error of estimate:

$$S^2_{Y.x_1,x_2} = \frac{\sum(Y - a - b_1 x_1 - b_2 x_2)^2}{n-3}$$

Substituting $\overline{Y}$ for the slope intercept term and rearranging:

$$S^2_{Y.x_1,x_2} = \frac{\sum\left[(Y-\overline{Y}) - b_1 x_1 - b_2 x_2\right]^2}{n-3}$$

Now, expressing $Y-\overline{Y}$ in deviation form and expanding the trinomial expression:

$$S^2_{Y.x_1,x_2} = \frac{1}{n-3}\sum\left[y^2 + b_1^2 x_1^2 + b_2^2 x_2^2 - 2yb_1 x_1 - 2yb_2 x_2 + 2b_1 b_2 x_1 x_2\right]$$

Bringing the summation operator inside and factoring constants:

$$S^2_{Y.x_1,x_2} = \frac{1}{n-3}\left(\sum y^2 + b_1^2 \sum x_1^2 + b_2^2 \sum x_2^2 - 2b_1 \sum x_1 y - 2b_2 \sum x_2 y + 2b_1 b_2 \sum x_1 x_2\right)$$

The following formulas can be used for simplification:

$$\sum y^2 = (n-1)S_y^2$$

$$\sum x_1^2 = (n-1)S_1^2$$

$$\sum x_2^2 = (n-1)S_2^2$$

$$\sum x_1 y = (n-1)r_{y1} S_y S_1$$

$$\sum x_2 y = (n-1)r_{y2} S_y S_2$$

$$\sum x_1 x_2 = (n-1)r_{12} S_1 S_2$$

For easy reference, these formulas are summarized in Table 2.

Table 2
Substitution Equations for Two Predictor Raw Score Model

$\sum y^2$	$= (n-1)S_y^2$
$\sum x_1^2$	$= (n-1)S_1^2$
$\sum x_2^2$	$= (n-1)S_2^2$
$\sum x_1 y$	$= (n-1)r_{y1} S_y S_1$
$\sum x_2 y$	$= (n-1)r_{y2} S_y S_2$
$\sum x_1 x_2$	$= (n-1)r_{12} S_1 S_2$

Note: equations are expressed in deviation score form. Each equation is based on algebraic rearrangements for basic sample descriptive statistics (compare Table 1). For example, the variance of Y is:

$$S_y^2 = \sum (Y-\bar{Y})^2/(n-1) = \sum y^2/(n-1)$$

Solving in terms of $\sum y^2$: $\sum y^2 = (n-1)S_y^2$.

Making these substitutions:

$$S^2_{Y.x_1,x_2} = \frac{1}{n-3}\left[(n-1)S_y^2 + (n-1)b_1^2 S_1^2 + (n-1)b_2^2 S_2^2 - 2(n-1)b_1 r_{y1} S_y S_1 - 2(n-1)b_2 r_{y2} S_y S_2 + 2(n-1)b_1 b_2 r_{12} S_1 S_2\right]$$

Factoring out the (n-1) term and rearranging:

$$S^2_{Y.x_1,x_2} = \frac{(n-1)}{(n-3)}\left[S_y^2 + (b_1^2 S_1^2 + b_2^2 S_2^2 + 2b_1 b_2 r_{12} S_1 S_2) - 2(b_1 r_{y1} S_y S_1 + b_2 r_{y2} S_y S_2)\right]$$

The next step is very important. The two terms in parentheses reduce to functions of the squared multiple R for two predictors. As was shown in the author's 1983a paper, the derivation of R for two predictors results in several equivalent ways to express R^2 or R^2. Table 3 shows forms of R^2 which will be used in the next step. (Compare O'Brien, 1983a, pages 12-18, especially p. 18).

Table 3

R^2 Functions of R^2 for Two Raw Score Predictors.[a]

$$R^2 = \frac{b_1^2 S_1^2 + b_2^2 S_2^2 + 2b_1 b_2 r_{12} S_1 S_2}{S_y^2} \cdot \frac{b_1 r_{y1} S_y S_1 + b_2 r_{y2} S_y S_2}{S_y^2}$$

Rearranging:

$$R^2 S_y^2 = b_1^2 S_1^2 + b_2^2 S_2^2 + 2b_1 b_2 r_{12} S_1 S_2 \cdot b_1 r_{y1} S_y S_1 + b_2 r_{y2} S_y S_2$$

Note: $R^2 = R^2_{Y.x_1,x_2}$

[a] See O'Brien, 1983a.

Thus,

$$R^2 S_y^2 = b_1^2 S_1^2 + b_2^2 S_2^2 + 2 b_1 b_2 r_{12} S_1 S_2$$

$$= b_1 r_{1y1} S_Y S_1 + b_2 r_{2y2} S_Y S_2$$

Making these substitutions:

$$S^2_{Y.x_1,x_2} = \frac{n-1}{n-3}\left[S_y^2 + S_y^2 R^2 - 2S_y^2 R\right]$$

$$= \frac{n-1}{n-3} S_y^2 \left[1 - R^2\right]$$

$$= \frac{n-1}{n-3} S_y^2 \left[1 - R^2_{Y.x_1,x_2}\right]$$

Taking the positive square root, the unbiased standard error of estimate for two raw score predictors is:

$$S_{Y.x_1,x_2} = S_y \sqrt{\frac{(n-1)}{(n-3)}\left[1 - R^2_{Y.x_1,x_2}\right]}$$

END OF PROOF

Derivation for Three Predictors

Prior to showing the derivation for the general case of p predictors, we will present the derivation for the three predictor model. This allows us to review the logic and procedures of the derivation. In addition, we introduce

summation notation throughout all of the steps of the derivation which simplifies the algebra for the general case.

For three raw score predictors, we will show that:

$$S_{Y.x_1,x_2,x_3} = S_y \sqrt{\frac{n-1}{n-4}\left[1 - R^2_{Y.x_1,x_2,x_3}\right]}$$

We begin by presenting the definition of the unbiased standard error of estimate for three predictors:

$$S_{Y.x_1,x_2,x_3} = \sqrt{\frac{\sum (Y - \hat{Y})^2}{n-(p+1)}}$$

$$= \sqrt{\frac{\sum (Y - a - b_1 x_1 - b_2 x_2 - b_3 x_3)^2}{n-4}}$$

As before, we will work with the variance error of estimate:

$$s^2_{Y.x_1,x_2,x_3} = \frac{\sum (Y - a - b_1 x_1 - b_2 x_2 - b_3 x_3)^2}{n-4}$$

Proceeding as before, we first replace a with $\bar{Y}$ and express $Y - \bar{Y}$ as y:

$$s^2_{Y.x_1,x_2,x_3} = \frac{\sum \left[(Y-\bar{Y}) - b_1 x_1 - b_2 x_2 - b_3 x_3\right]^2}{n-4}$$

$$= \frac{\sum \left[y - b_1 x_1 - b_2 x_2 - b_3 x_3\right]^2}{n-4}$$

Expanding this quadrinomial expression:

$$s^2_{Y.x_1,x_2,x_3} = \frac{1}{n-4} \sum \left[y^2 + b_1^2 x_1^2 + b_2^2 x_2^2 + b_3^2 x_3^2 - 2yb_1 x_1 - 2yb_2 x_2 - 2yb_3 x_3 + 2b_1 b_2 x_1 x_2 + 2b_1 b_3 x_1 x_3 + 2b_2 b_3 x_2 x_3 \right]$$

Bringing the summation operator inside:

$$S^2_{Y.x_1,x_2,x_3} = \frac{1}{n-4}\left[\sum y^2 + b_1^2\sum x_1^2 + b_2^2\sum x_2^2 + b_3^2\sum x_3^2 - 2b_1\sum x_1 y - 2b_2\sum x_2 y - 2b_3\sum x_3 y + 2b_1 b_2\sum x_1 x_2 + 2b_1 b_3\sum x_1 x_3 + 2b_2 b_3\sum x_2 x_3\right]$$

The following substitution formulas stated in general form will help us to simplify the above expression (see Table 4 for reference):

$$S^2_y = (n-1)\sum y^2$$

For any x_j:

$$S^2_j = (n-1)\sum x_j^2$$

For any $x_Y x_j$

$$\sum x_Y x_j = (n-1) r_{yj} S_Y S_j$$

For any $x_i x_j$:

$$\sum x_i x_j = (n-1) r_{ij} S_i S_j$$

Applying these substitutions:

$$S^{2}_{Y.x_1,x_2,x_3} = \frac{1}{n-4}\left[(n-1)S^{2}_{y}\right.$$

$$+ (n-1)\,b_1\,S^{2}_{1} + (n-1)\,b_2\,S^{2}_{2} + (n-1)\,b_3\,S^{2}_{3}$$

$$- 2(n-1)\,b_1\,r_{1y_1}\,S_1\,S_{y_1} - 2(n-1)\,b_2\,r_{2y_2}\,S_2\,S_{y_2} - 2(n-1)\,b_3\,r_{3y_3}\,S_3\,S_{y_3}$$

$$\left. + 2(n-1)\,b_1\,b_2\,r_{12}\,S_1\,S_2 + 2(n-1)\,b_1\,b_3\,r_{13}\,S_1\,S_3 + 2(n-1)\,b_2\,b_3\,r_{23}\,S_2\,S_3\right]$$

Table 4
Generalized Substitution Equations For Raw Score Model[a]

S_y^2	•	$(n-1)\sum y^2$
S_j^2	•	$(n-1)\sum x_j^2$
$\sum x_{y.} x_{.j}$	•	$(n-1) r_{yj} S_y S_j$
$\sum x_i x_j$	•	$(n-1) r_{ij} S_i S_j$

[a] For example, the second equation applies to any X variable; for the jth X variable, the sum of squares is related to the jth variance.

Factoring out (n-1) and rearranging:

$$S^2_{Y.x_1,x_2,x_3} = \frac{n-1}{n-4}\left[S^2_y + (b_1^2 S_1^2 + b_2^2 S_2^2 + b_3^2 S_3^2 + 2b_1 b_2 r_{12} S_1 S_2 + 2b_1 b_3 r_{13} S_1 S_3 + 2b_2 b_3 r_{23} S_2 S_3) - 2(b_1 r_{y1} S_1 S_y + b_2 r_{y2} S_2 S_y + b_3 r_{y3} S_3 S_y)\right]$$

We now express the parenthesized terms in summation notation (see O'Brien, 1983a):

$$S^2_{Y.x_1,x_2,x_3} = \frac{n-1}{n-4}\left[S^2_y + \left(\sum_{j=1}^{3} b_j^2 S_j^2 + 2\sum_{j=2}^{3}\sum_{i=1}^{2} b_i b_j r_{ij} S_i S_j\right) - 2\left(\sum_{j=1}^{3} b_j r_{jy} S_j S_Y\right)\right]$$

Table 5 shows equivalent forms of R^2 for three predictors stated in summation notation.

Table 5

Functions of R^2 For Three Raw Score Predictors[a]

$$R^2 = \frac{\sum_{j=1}^{3} b_j^2 S_j^2 + 2\sum_{j=2}^{3}\sum_{i=1}^{2} b_i b_j r_{ij} S_i S_j}{S_y^2} \cdot \frac{\sum_{j=1}^{3} b_j r_{yj} S_Y S_j}{S_y^2}$$

Rearranging:

$$R^2 S_y^2 = \sum_{j=1}^{3} b_j^2 S_j^2 + 2\sum_{j=2}^{3}\sum_{i=1}^{2} b_i b_j r_{ij} S_i S_j \cdot \sum_{j=1}^{3} b_j r_{yj} S_y S_j$$

Note: $R^2 = R^2_{Y.x_1,x_2,x_2}$

[a] See O'Brien, 1983a

Thus:

$$R^2 S_y^2 = \sum_{j=1}^{3} b_j^2 S_j^2 + 2 \sum_{j=2}^{3} \sum_{i=1}^{2} b_i b_j r_{ij} S_i S_j$$

$$= \sum_{j=1}^{3} b_j r_{yj} S_y S_j$$

Substituting:

$$S^2_{Y.x_1,x_2,x_3} = \frac{n-1}{n-4} \left[S_y^2 + S_y^2 R^2 - 2 S_y^2 R^2 \right]$$

Simplifying:

$$S^2_{Y.x_1,x_2,x_3} = \frac{n-1}{n-4} S_y^2 \left[1 - R^2\right]$$

or

$$S^2_{Y.x_1,x_2,x_3} = \frac{n-1}{n-4} S_y^2 \left[1 - R^2_{Y.x_1,x_2,x_3}\right]$$

Therefore, the unbiased standard error of estimate is:

$$S_{Y.x_1,x_2,x_3} = S_y \sqrt{\frac{n-1}{n-4}\left[1 - R^2_{Y.x_1,x_2,x_3}\right]}$$ END OF PROOF

Derivation For p Predictors

In this section, we show the general form of the unbiased standard error of estimate when the regression model contains some unknown but finite number of predictors (p).
We will follow the same steps in the derivation we used for one, two and three predictors. It will be seen that the derivation for the general case of p predictors is a straightforward multivariate generalization.

Formally, we will show that the unbiased standard error of estimate for p predictors is:

$$S^2_{Y.x_1,x_2,\ldots,x_j,\ldots,x_p} = S_y \sqrt{\frac{n-1}{n-(p+1)}\left[1 - R^2_{Y.x_1,x_2,\ldots,x_j,\ldots,x_p}\right]}$$

Definitions for terms in the formula were given in the section "Overview of Derivation".

Starting with the definition of the unbiased standard error of estimate:

$$S_{Y.x_1,x_2,\ldots x_j,\ldots,x_p} = \sqrt{\frac{\sum (Y - \hat{Y})^2}{n-(p+1)}}$$

$$= \sqrt{\frac{\sum \left[Y - a - b_1x_1 - b_2x_2 - \ldots - b_jx_j - \ldots - b_px_p\right]^2}{n-(p+1)}}$$

As in the previous derivations, we will work with the variance error of estimate:

$$S^2_{Y.x_1,x_2,\ldots,x_j,\ldots,x_p} = \frac{\sum (Y - a - b_1x_1 - b_2x_2 - \ldots - b_jx_j - \ldots - b_px_p)^2}{n-(p+1)}$$

Now replace a by $\bar{Y}$, and express $Y-\bar{Y}$ in deviation score form:

$$S^2_{Y.x_1,x_2,\ldots,x_j,\ldots,x_p} = \frac{\sum (y - b_1x_1 - b_2x_2 - \ldots - b_jx_j - \ldots - b_px_p)^2}{n-(p+1)}$$

Expanding this multinomial:

$$S^2_{Y.x_1,x_2,\ldots,x_j,\ldots,x_p} \cdot \frac{1}{n-(p+1)} X$$

$$\sum (y^2 + b_1^2 x_1^2 x_1^2 + b_2^2 x_2^2 x_2^2 + \ldots + b_j^2 x_j^2 x_j^2 + \ldots + b_p^2 x_p^2 x_p^2$$

$$- 2yb_1 x_1 x_1 - 2yb_2 x_2 x_2 - \ldots - 2yb_j x_j x_j - \ldots - 2yb_p x_p x_p$$

$$+ 2b_1 b_2 x_1 x_2 x_1 x_2 + 2b_1 b_3 x_1 x_3 x_1 x_3 + \ldots + 2b_i b_j x_i x_j x_i x_j + \ldots + 2b_{p-1} b_p x_{p-1} x_p x_{p-1} x_p)$$

Bringing the summation operator inside:

$$S^2_{Y.x_1,x_2,\ldots,x_j,\ldots,x_p} \cdot \frac{1}{n-(p+1)} X$$

$$(\sum y^2 + b_1^2 \sum x_1^2 x_1^2 + b_2^2 \sum x_2^2 x_2^2 + \ldots + b_j^2 \sum x_j^2 x_j^2 + \ldots + b_p^2 \sum x_p^2 x_p^2$$

$$- 2b_1 \sum yx_1 x_1 - 2b_2 \sum yx_2 x_2 - \ldots - 2b_j \sum yx_j x_j - \ldots - 2b_p \sum yx_p x_p$$

$$+ 2b_1 b_2 \sum x_1 x_2 x_1 x_2 + 2b_1 b_3 \sum x_1 x_3 x_1 x_3 + \ldots + 2b_i b_j \sum x_i x_j x_i x_j + \ldots + 2b_{p-1} b_p \sum x_{p-1} x_p x_{p-1} x_p)$$

Using the generalized substitution formulas given in Table 4, we can simplify as follows:

$$S^{2}_{Y.x_1,x_2,\ldots,x_j,\ldots,x_p} \cdot \frac{1}{n-(p+1)} X$$

$$\Big[(n-1)S^{2}_{y} + (n-1)b_{1\,1}\,S^{2}_{2} + (n-1)b_{2\,2}\,S^{2}_{2} + \ldots + (n-1)b_{p\,p}\,S^{2}_{p} \Big]$$

$$- 2(n-1)b_{1\,y1\,y\,1}\,r\,S\,S - 2(n-1)b_{2\,y2\,y\,2}\,r\,S\,S - \ldots - 2(n-1)b_{j\,yj\,y\,j}\,r\,S\,S - \ldots -$$

$$2(n-1)b_{p\,yp\,y\,p}\,r\,S\,S$$

$$+ 2(n-1)b_{1\,2\,12\,1\,2}\,b\,r\,S\,S + 2(n-1)b_{1\,3\,13\,1\,3}\,b\,r\,S\,S + \ldots + 2(n-1)b_{i\,j\,ij\,i\,j}\,b\,r\,S\,S + \ldots +$$

$$\Big[2(n-1)b_{p-1\,p\,p-1,p\,p-1\,p}\,b\,r\,S\,S \Big]$$

Factoring out (n-1) and rearranging:

$$S^{2}_{Y.x_1,x_2,\ldots,x_j,\ldots,x_p} \cdot \frac{n-1}{n-(p+1)} X$$

$$\Big[S^{2}_{y} + (b_{1\,1}\,S^{2}_{2} + b_{2\,2}\,S^{2}_{2} + \ldots + b_{j\,j}\,S^{2}_{2} + \ldots + b_{p\,p}\,S^{2}_{2} +$$

$$2b_{1\,2\,12\,1\,2}\,b\,r\,S\,S + 2b_{1\,3\,13\,1\,3}\,b\,r\,S\,S + \ldots + 2b_{i\,j\,ij\,i\,j}\,b\,r\,S\,S + \ldots +$$

$$2b_{p-1\,p\,p-1,p\,p-1\,p}\,b\,r\,S\,S \,)$$

$$- 2(b_{1\,y1\,y\,1}\,r\,S\,S + b_{2\,y2\,y\,2}\,r\,S\,S + \ldots + b_{j\,yj\,y\,j}\,r\,S\,S + \ldots + b_{p\,yp\,y\,p}\,r\,S\,S \,) \Big]$$

Expressing the terms in parentheses in summation notation:

$$S^2_{Y.x_1,x_2,\ldots,x_j,\ldots,x_p} = \frac{n-1}{n-(p+1)} X$$

$$\left[S^2_y + \left(\sum_{j=1}^{p} b_j^2 S_j^2 + 2 \sum_{j=2}^{p} \sum_{i=1}^{p-1} b_i b_j r_{ij} S_i S_j \right) - 2 \left(\sum_{j=1}^{p} b_j r_{yj} S_y S_j \right) \right]$$

Table 6 shows equivalent forms of the multiple R^2 for p predictors (see O'Brien, 1983a).

Table 6

Functions of R^2 for p Predictors[a]

$$R^2 = \frac{\sum_{j=1}^{p} b_j^2 S_j^2 + 2\sum_{j=2}^{p}\sum_{i=1}^{p-1} b_i b_j r_{ij} S_i S_j}{S_y^2} \cdot \frac{\sum_{j=1}^{p} b_j r_{yj} S_Y S_j}{S_y^2}$$

Rearranging:

$$R^2 S_y^2 = \sum_{j=1}^{p} b_j^2 S_j^2 + 2\sum_{j=2}^{p}\sum_{i=1}^{p-1} b_i b_j r_{ij} S_i S_j \cdot \sum_{j=1}^{p} b_j r_{yj} S_y S_j$$

Note: $R^2 = R^2_{Y.x_1, x_2, \ldots, x_j, \ldots, x_p}$

[a] See O'Brien, 1983a

Thus:

$$R^2 S_y^2 = \sum_{j=1}^{p} b_j^2 S_j^2 + 2 \sum_{j=2}^{p} \sum_{i=1}^{p-1} b_i b_j r_{ij} S_i S_j$$

$$= \sum_{j=1}^{p} b_j r_{yj} S_y S_j$$

Substituting into the variance error of estimate above:

$$S^2_{Y.x_1,x_2,\ldots,x_j,\ldots,x_p} = \frac{n-1}{n-(p+1)} \left[S_y^2 + R^2 S_y^2 - 2R^2 S_y^2 \right]$$

$$S^2_{Y.x_1,x_2,\ldots,x_j,\ldots,x_p} = \frac{n-1}{n-(p+1)} S_y^2 \left[1 - R^2_{y.x_1,x_2,\ldots,x_j,\ldots,x_p} \right]$$

Therefore:

$$S_{Y.x_1,x_2,\ldots,x_j,\ldots,x_p} = S_y \sqrt{\frac{n-1}{n-(p+1)} \left[1 - R^2_{Y.x_1,x_2,\ldots,x_j,\ldots,x_p} \right]}$$

END OF PROOF

Derivations for Standard Score Model

Introduction

We have presented derivations for the unbiased standard error of estimate for the linear raw score model when the number of predictors was one, two, three and some finite number, p. In this part of the paper we will outline the derivations for the standard score model.

The reader may be aware of the fact that there is a simple relationship between models in raw score form and standard score (Z) form. This relationship obviates the need for presenting detailed derivations for the Z score model. Therefore, we will outline the derivations for the standard score model, and leave the proofs as an exercise for the reader. We will show the logic behind transforming from the linear raw score model to the Z score model. First we take the standardized model for one predictor. We then provide an outline for generalizing the derivation for the p predictor standard score case.

Derivation for One Predictor

Recall the derivation for the one predictor raw score model. The derivation of the standard error of estimate was shown to be:

$$S_{Y.x_1} = S_y \sqrt{\frac{n-1}{n-2}\left[1-r^2_{x_1 y}\right]}$$

Let us now consider the model in standard score form. First, recall the following relationships for the Z score model (See O'Brien, 1982b for proofs):

$$S_{Z_Y} = 1$$

$$r^2_{Z_Y, Z_{y1}} = r^2_{y1}$$

That is, the standard deviation for the raw score variable Y is equal to unity when Y is standardized. Also, the square of the simple (zero order) Pearson correlation when calculated in raw score form is identical to the correlation between the same variables that have each been standardized. Taking these facts into account, we can rewrite the raw score standard error of estimate for Z scores as follows:

$$S_{Z_Y . Z_1} = \sqrt{\frac{\sum (Z_Y - Z_{\hat{Y}})^2}{n-2}}$$

$$= S_{Z_y} \sqrt{\frac{n-1}{n-2}\left[1 - r^2_{Z_Y, Z_1}\right]}$$

$$= 1 \sqrt{\frac{n-1}{n-2}\left[1 - r^2_{x\,y}\right]}$$

$$= \sqrt{\frac{n-1}{n-2}\left[1 - r^2_{x\,y}\right]}$$

$$= S_{Y.x_1}$$

If one were to extend this logic to the case of p standardized predictors, the standard error of estimate for p standardized predictors is:

$$S_{Z_Y.Z_1,Z_2,\ldots,Z_j,\ldots Z_p} = S_{Z_y}\sqrt{\frac{n-1}{n-(p+1)}\left[1 - R^2_{Z_Y.Z_1,Z_2,\ldots,Z_j,\ldots,Z_p}\right]}$$

$$\cdot \sqrt{\frac{(n-1)\left[1- R^2_{Y.x_1,x_2,\ldots,x_j,\ldots,x_p}\right]}{n-(p+1)}}$$

For the p predictor case S_{Z_y} also is equal to 1.

It remains to be proved that the squared multiple R's are equal to one another. It can be shown that they are equal for p predictors, although this statement is not proved in this paper.

Outline for Derivations

The reader who desires to derive the unbiased standard error of estimate for p linear standardized predictors may use the following outline as a guide. Essentially, the steps parallel those for the raw score model. First, the definitional form for the standard error of estimate is stated. Substituting the terms of the regression model for p predictors is the second step. (See O'Brien, 1983c). Third, square the multinomial expression. Next, a series of equations are substituted into the squares and cross products of the squared multinomial. The reader may refer to the author's paper (1983c) for the relevant equations. The simplified expression is then expressed in summation notation. Functions of the multiple squared R are substituted. Upon simplification, the result will be the unbiased standard error of estimate for the Z score model.

Many students who work out the derivations for the Z score model prefer to work with several predictors in succession. This was our approach for the raw score model derivations. A careful review of the steps used in the raw score derivations may be helpful in working through the long tedious algebra.

9

Appendix A

Errata for "A derivation of the sample multiple correlation formula for raw scores, ED 235 205

Page*	Now Reads	Correct to
10, footnote, 3 lines down	$\bar{X}\ \bar{Y}$	$\bar{X}\ \bar{Y}$ 1
10, footnote, 4 lines down	$n\ \bar{X}\ \bar{Y}$	$n\ \bar{X}\ \bar{Y}$ 1
13	$\text{var}(b_2, x_2)$	$\text{var}(b_2 x_2)$
16, footnote, last 2 lines	... and simplifying. See text for details.	See the text for details.
17, footnote	Multiple R	multiple R
24, footnote 1	$i \neq j$	Omit this.
29, equation $x_p\ Y_p$	$b_p \sum x_p$	$b_p \sum x_p^2$
30, 3 lines from bottom	$b_{2j}\ b_{2j}\ r_j\ S_{2j}\ S_j$	$b_{2p}\ b_{2p}\ r_p\ S_{2p}\ S_p$
34, 2nd equation	$= \ldots + b_{yj}\ r_j\ S_{yj}\ S_j$	change = to +
36, 2 lines from bottom of text	mathematical calculus	mathematical statistics
38, 2nd equation	S_1^2	S_y^2
43, last line in text	$S_j^2 = o$	$S_j^2 = 1$

* Page number at top of text.

Appendix B

Discussion of Linear and Nonlinear Regression Models

This appendix will clarify terminology used in two previous papers (O'Brien, 1982c, 1983a). Some readers have requested clarification of my use of terms "linear" and "nonlinear' as they apply to regression analysis.

There are two reasons why this should be done. First, the terminology and/or notation used in applied social science statistics textbooks and similar sources is quite variable. This has the potential for causing confusion in students' minds when attempting to read the same subject matter in different sources. Second, it is very important to be clear about the differences between a linear and nonlinear regression model. As will be seen, "truly" nonlinear regression models are not often used in many areas of social science.

Our aim in this appendix merely is to clarify the uses of the terminology. References are cited at the end of the appendix for readers who desire to learn more about nonlinear regression models.

I believe confusion exists in the use of the terminology for several reasons. Perhaps the basic factor relates to what students learn in nonstatistical mathematical courses. The terms linear/nonlinear as they relate to functions or relationships discussed in mathematics textbooks are not used in the same way by statisticians when discussing linear/nonlinear regression models.

Consider a simple example of the parabola (or quadratic or second degree equation):

$$Y = f(X) = 9 - X^2 \qquad -3 < X < 3$$

If this function is plotted on ordinary graphing paper for values of X $\pm$ 3 , the plot would show a curve opening downward

with maximum height of 9 Y units at the origin. This function is not linear in form because it cannot be expressed in the form of a first degree equation:

$$Y = f(X) = a + bX$$

Geometrically, a plot of the quadratic function above would not reveal a straight line or linear function. For these two reasons, the parabola may be thought of as a "nonlinear" function.

Statisticians use the terms linear/nonlinear in a different manner. In the statisticians use of the terms, the difference between them has more to do with the form of the regression parameters (slope terms) than with the form of the independent or dependent variables. In addition, a plot of the raw observed data points is not relevant to classifying a regression model as linear or nonlinear.

Let us examine some examples. Assume the following regression model (adapted from Draper and Smith, p. 264):

$$F = \exp(b_1 + b_2 X^2 + e) \qquad (1)$$

Where:

- F = the dependent variable,
- exp = the exponentiation operator for the mathematical constant, e = 2.71828 (approx.),
- b_1, b_2 = parameters to be estimated,
- X_1 = the independent variable,
- e = the stochastic error term (as used in this paper).

Note that equation 1 expresses what we have been calling a "raw score model"; e.g., for equation 1, we could write:

$$F = \exp(\hat{F} + e).$$

Is the model in (1) a linear or nonlinear regression model? We need to examine the terms in (1) to decide.

Let us now rework equation 1 to render the model linear. If we take the natural logarithm of each side of equation 1, we obtain:

$$\ln F = \ln\left[\exp(b_1 + b_2 X^2 + e)\right]$$

$$= b_1 + b_2 X^2 + e \qquad (2)$$

We now redefine the terms in equation 2. Let:

$$Y = \ln F,$$
$$X = X^2$$

Then (2) becomes:

$$Y = b_1 + b_2 X + e \qquad (3)$$

Equation 2 has been linearized. Statisticians would call the regression model expressed in (3) a linear model despite the fact that the relationship between the dependent and independent variables is not one of a straight line.

Draper and Smith offer useful terminology to distinguish (1) from (3). The regression model stated in (1) may be referred to as intrinsically linear. This means that although equation 1 is nonlinear (with respect to the parameters b_1 and b_2), transformations can be made to express the model in a form which is linear (with respect to the parameters).

To take a second example (also from Draper and Smith), consider the following regression model:

$$G = \frac{b_1^{b_2}}{b_1 - b_2}\left[\exp(-b_2 X) - \exp(-b_1 X)\right] + e \qquad (4)$$

Where:

G = the dependent variable,
exp = as in equation 1,
b_1, b_2 = the parameters,

X = the independent variable

This model is nonlinear (with respect to the parameters). In addition, equation 4 cannot be transformed such that the parameters will be linear in

form. Draper and Smith refer to such a regression model as intrinsically nonlinear.

Further discussion and examples of linear/nonlinear regression models may be found in Kendall and Stuart (1967), Mosteller and Tukey (1977) and Nie, et al. (1975). Those references provide additional source material.

Notes

1
See O'Brien (1983a, Appendix B) for an errata sheet. Page references given in the errata pertain to the original pagination (i.e., at the top of the page).

2
Errata for this paper are given in Appendix A of the present paper.

3
Readers who need to review regression analysis theory can refer to standard applied statistics textbooks. One that is highly recommended for its thoroughness and clarity is by Lindeman, Gold and Merenda (1982). A general overview is given by Lewis-Beck (1980).

4
See Appendix B for discussion of linear and nonlinear regression models.

5
If it is understood that the summation limits range from the first observation (i=1) to the last (i=n), then we can drop the summation limits; n refers to the total number of observations for the criterion and predictor(s). This sample size is the same regardless of the number of predictors. Later when the algebra becomes more complex, we use summation limits extensively.

6
As mentioned earlier, it is assumed that the reader is familiar with the author's 1983a paper.

7
The regression model for one standardized predictor is:

$$Z_{\hat{Y}} = A + B_1 Z_1$$

The observed standard score model is:

$$Z_Y = Z_{\hat{Y}} + e_Z$$

where:

$Z_{\hat{Y}}$ = the predicted criterion in standard score form,

A = the slope intercept term (not standardized--see O'Brien, 1982c)

Z_1 = the standardized predictor; i.e.,

$$Z_1 = (X_1 - \bar{X}_1)/S_1 \quad \text{where } S_1 \text{ is the standard deviation of } X_1$$

B_1 = slope term (regression or beta weight)

e_Z = the prediction error.

8

The reader may wonder why we divide by the term, n-2. This term represents the "degrees of freedom" for the unbiased standard error of estimate for one predictor.
It can be shown that dividing by the appropraite degrees of freedom term makes the sample standard error of estimate unbiased; i.e., the expected value of the sample standard error of estimate equals the population parameter.
In general, the degrees of freedom for the unbiased standard error of estimate is: n-(p+1) , where p = the number of predictors in the regression model. For one predictor, n-(p+1) = n-(1+1) = n-2. p + 1 arises from the number of parameters that can be estimated in any raw score linear regression model--p slope (b) terms plus the slope intercept term. For a good discussion of degrees of freedom, see the classic paper by Helen Walker (1940, 1971). See also Stilson (1966).

9

An alternate approach to the derivations could be used by working with matrix albegra notation. The author intends to present the derivations of this paper and others in this series in matrix algebra. They will be written as part of this series for ERIC.

References

Draper, Norman and Harry Smith. Applied Regression Analysis. New York: John Wiley & Sons, 1966.

Kendall, Maurice G. and Alan Stuart. The Advanced Theory of Statistics: Inference and Relationship, Vol. 2. (2nd ed.). New York: Hafner Publishing Co., 1967.

Lewis-Beck, Michael S. Applied Regression: An Introduction. Beverly Hills, CA: Sage Publication, 1980.

Lindeman, Richard H., Ruth Gold and Peter Merenda. Bivariate and Multivariate Analysis. Chicago: Scott, Foresman and Co., 1982

Mosteller, Frederick and John W. Tukey. Data Analysis and Regression: A Second Course in Statistics. Mass: Addison-Wesley Publishing Co., 1977.

Nie, Norman H. et al. Statistical Package for the Social Sciences (2nd ed.). NY: McGraw Hill Book Co., 1975

O'Brien, Francis J., Jr. A proof that t^2 and F are identical: the general case, 1982a. ERIC ED 215 894

_____. Proof that the sample bivariate correlation coefficient has limits ± 1, 1982b. ERIC ED 216 874

_____. A derivation of the sample multiple correlation formula for standard scores, 1982c. ERIC ED 223 429

_____. A derivation of the sample multiple correlation formula for raw scores, 1983a. ERIC ED 235 205

Stilson, Donald W. Probability and Statistics in Psychological Research and Theory. San Francisco, CA: Holden-Day, Inc., 1966.

Walker, Helen A. Degrees of Freedom. Journal of Educational Psychology, 31, 1940, 239-269. Reprinted in Readings in Statistics for the Behavioral Scientist. Joseph A. Steger (Ed.). NY: Holt, Rinehart & Winston, Inc., 1971.